Gain Greater Productivity in Enterprise Data Mining

Discover patterns and associations in your organization's data and make decisions that lead to significant, measurable improvements in results with Clementine from SPSS Inc. Clementine is the leading data mining workbench, popular worldwide with data miners and business analysts alike, and uniquely supports the entire data mining process. With Clementine and its associated SPSS products, users can easily access and prepare numeric, text, and Web data for modeling; rapidly build and compare models; and efficiently deploy them, in real time, to people and systems making decisions and recommendations.

Because it enables you to seamlessly integrate data mining results with other business systems and processes, Clementine helps your organization make faster, better decisions, enterprise wide.

Data miners can pursue natural "train of thought" analysis, thanks to Clementine's powerful visual interface. Clementine produces "streams," visual maps of each step in the data mining process. By interacting with a stream, analysts can add business knowledge—the key to successful data mining—at any point in the process. Because the interface allows analysts to focus on knowledge discovery rather than on performing technical tasks such as writing code, your data mining process is far more efficient.

With Clementine 9.0, you can improve data mining productivity enterprise wide, thanks to SPSS Model Manager™ capabilities. Fully integrated with Clementine, Model Manager enables you to leverage your entire organization's business knowledge. Data miners save streams, models, and output files in a central, searchable repository. Authorized users can then access and reuse the most effective streams and models.

Administrators can easily ensure that only those with the proper authority can see and interact with streams and models. Model Manager also includes a versioning feature, so you can be confident that production versions don't get overwritten. With Model Manager, you have complete control of your data mining assets.

Select from an unparalleled breadth of techniques

Clementine includes a host of analytical techniques for obtaining useful, reliable data mining results. SPSS Inc. has more than 35 years' experience in predictive analytics, and our algorithms are calibrated and verified to support the creation of powerful data mining models. Clementine has consistently offered a broader range of machine-learning and statistical techniques than any other data mining workbench. You have a choice of algorithms for clustering, classification, association, and prediction.

Integrate with existing information systems

Clementine is an open, standards-based solution that easily integrates with your organization's existing IT systems. Because it efficiently delivers information to support decision making at all levels of your organization, Clementine helps your IT department meet internal customer needs while helping your organization to gain even greater value from your existing technology investments.

Deliver data mining results efficiently

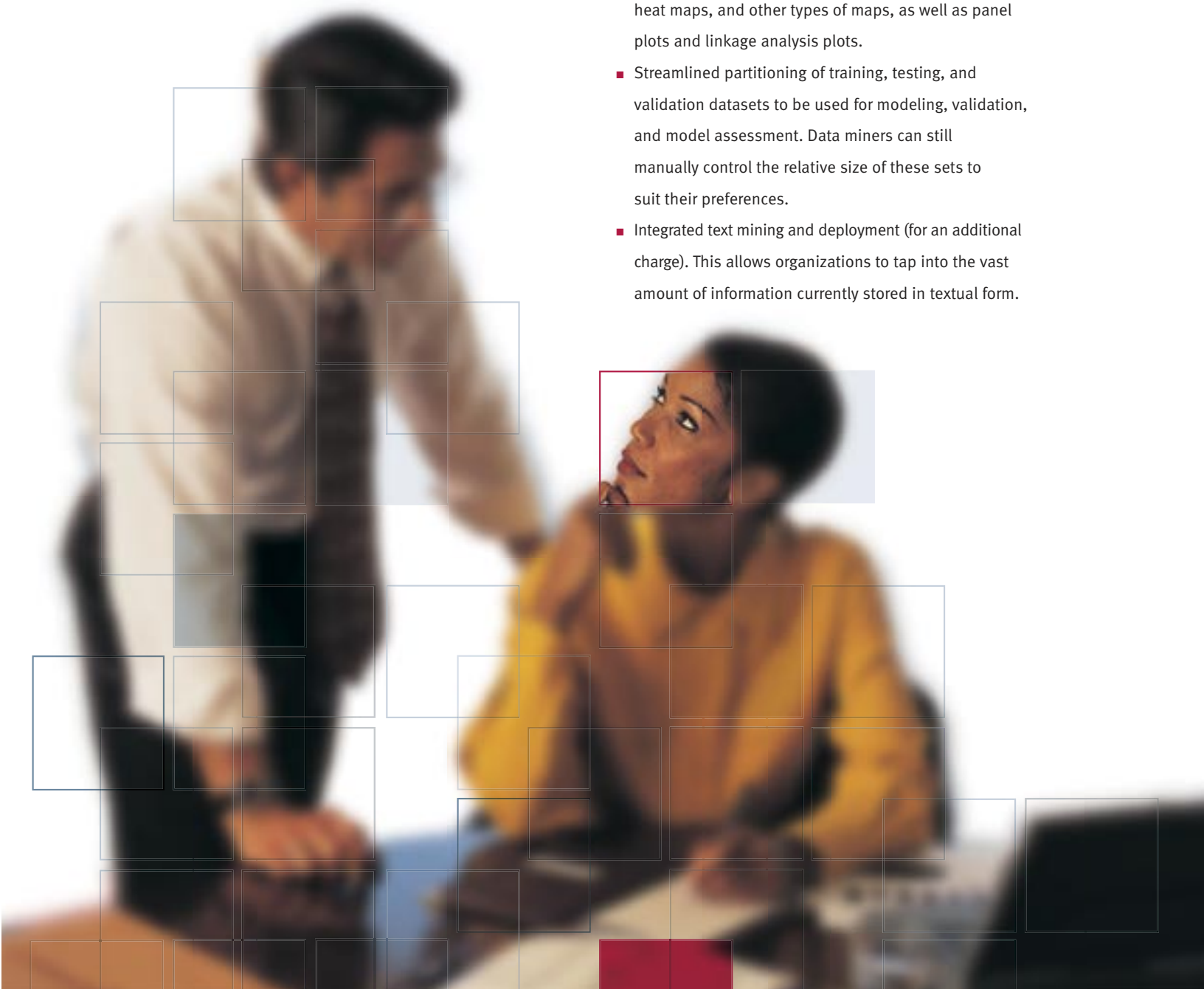
SPSS offers a number of deployment options to meet your needs for in-database or real-time scoring. Clementine exports not only the model but also all data mining steps—including data access, modeling, and post-processing—in industry-standard Predictive Model Markup Language (PMML). This saves your organization time when deploying to operational systems—so you see the positive effects of predictive analytics sooner.

Major enhancements in Clementine 9.0

With this release, SPSS continues its commitment to delivering an unmatched breadth of analytic techniques, built on an open architecture that supports greater flexibility in modeling and deployment and a higher return on your database and data mining assets.

This release includes:

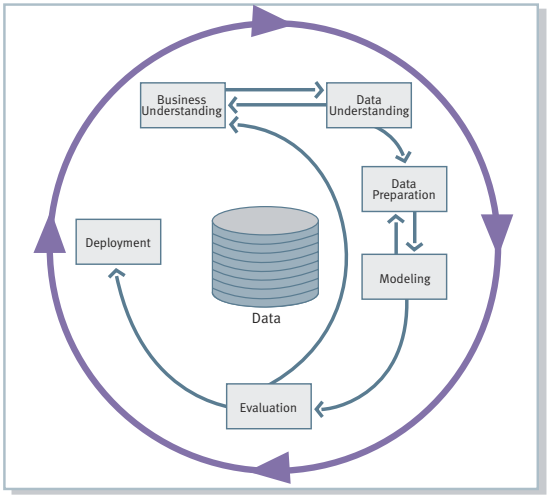
- Integration with SPSS Model Manager. This provides centralized control of data mining assets for improved data mining productivity enterprise wide.
- Enhanced in-database mining and modeling. Organizations can improve the speed and efficiency with which they conduct data mining within IBM®, Oracle®, or Microsoft® databases.
- Support for interactive building and user-defined splitting of decision tree models. This is accomplished through new modeling algorithms: CHAID, Exhaustive CHAID, and QUEST, as well as support for C&RT.
- Improved visualization. Using the included Advanced Visualization for Clementine add-on module, data miners can create bar charts, pie charts, boxplots, scatterplot matrices (SPLOM), parallel coordinate maps, heat maps, and other types of maps, as well as panel plots and linkage analysis plots.
- Streamlined partitioning of training, testing, and validation datasets to be used for modeling, validation, and model assessment. Data miners can still manually control the relative size of these sets to suit their preferences.
- Integrated text mining and deployment (for an additional charge). This allows organizations to tap into the vast amount of information currently stored in textual form.



Features*

CRISP-DM

With Clementine, your company's data miners can focus on business problem solving, rather than on programming. At every step, Clementine supports the de facto industry standard, the CRoss-Industry Standard Process for Data Mining (CRISP–DM). Clementine projects can be efficiently organized using the CRISP–DM project manager. And, thanks to SPSS Model Manager, you can support more efficient data mining enterprise wide.



The CRISP–DM process, as shown above, enables data miners to efficiently implement data mining projects that yield measurable business results.

Business understanding

Clementine's visual interface makes it easy for your data miners to apply business knowledge to data mining projects. In addition, optional business-specific Clementine Application Templates (CATs) are available to help you get results faster. CATs ship with sample data that can be installed as flat files or as tables in a relational database schema. CATs include the:

- CRM CAT**
- Telco CAT**
- Fraud CAT**
- Microarray CAT**
- Web Mining CAT** (requires the purchase of Web Mining for Clementine™)

Data understanding

- ▣ Enjoy new graph types with Advanced Visualization for Clementine, an included add-on. Create bar charts, pie charts, boxplots, SPLOM, heat maps, parallel coordinate maps, panel plots, and linkage analysis plots.
- Obtain a comprehensive first look at your data by using Clementine's data audit node
- Visually interact with your data
  - Select a region of a graphic and view the selected information in a table, or use this information downstream
- Create histograms, distributions, line plots, and point plots
- Use Web association detection
- Display 3-D, panel, and animated graphs
- View data quickly through graphs, summary statistics, or an assessment of data quality

Data preparation

- Access data
  - Structured (tabular) data
    - Access ODBC-compliant data sources with the SPSS Data Access Pack, which ships with Clementine. Drivers in this middleware pack include support for IBM DB2®, Oracle, Microsoft SQL Server™, Informix®, and Sybase® databases.
    - Import delimited and fixed-width text files; any SPSS® file; and SAS® 6,7,8, and 9 files
  - Unstructured (textual) data
    - Automatically extract concepts from documents and from text notes in databases using Text Mining for Clementine**
  - Web site data
    - Automatically extract Web site events from Web logs using Web Mining for Clementine**
  - Data output
    - Work with delimited and fixed-width text files; ODBC; Microsoft Excel™; SPSS; and SAS 6,7,8, and 9 files
- Choose from various data-cleaning options
  - Remove or replace invalid data
  - Automatically fill in missing values
- Manipulate data
  - ▣ Partition data into training, test, and validation datasets
  - Work with complete record and field operations, including:
    - Field filtering, naming, derivation, binning, re-categorization, value replacement, and field reordering
    - Record selection, sampling, merging (through inner joins, full outer joins, partial outer joins, and anti-joins), and concatenation; sorting, aggregation, and balancing; deriving new fields based on conditional criteria; and calculating new fields
    - Specialized manipulations for showing the "history" of values and converting set variables into flag variables

Modeling

- Mine data in the database where it resides with in-database modeling. Support:
  - IBM DB2 Enterprise Edition 8.2 decision trees, regression, association, and demographic clustering techniques
  - Oracle 10*g* Naïve Bayes and Adaptive Bayes networks and Support Vector Machines (SVM)
  - Microsoft SQL Server 2000 Analysis Services decision trees
- Use predictive and classification techniques
  - Neural networks (multi-layer perceptrons using error back-propagation, radial basis function, and Kohonen networks)
- Browse the importance of the predictors
  - Decision trees and rule induction techniques, including CHAID, exhaustive CHAID, QUEST, and C&RT
  - Browse and interactively create splits in decision trees
  - Rule induction techniques in C5.0
    - Browse, collapse, and expand decision rules
  - Linear regression, logistic regression, and multinomial logistic regression
    - View model equations and advanced statistical output
- Use clustering and segmentation techniques
  - Kohonen networks, K-means, and TwoStep
    - View cluster characteristics with a graphical viewer

** Features are subject to change based on the final product release* ☐ *Symbol indicates a new feature* ***Separately priced modules*

- Choose from several association detection algorithms
  - GRI, Apriori, sequence, and CARMA algorithms
    - Score data using models generated by association detection algorithms
    - Filter, sort, and create subsets of association models using the association model viewer
- Employ data reduction techniques
  - Factor analysis and principal components analysis
    - View model equation and advanced statistical output
- Combine models through meta-modeling
  - Multiple models can be combined, or one model can be used to build a second model
- Import PMML-generated models created in other tools such as AnswerTree® and SPSS
- Use Clementine External Module Interfaces (CEMI) for custom algorithms
  - Purchase add-on tools from the Clementine Plus Program
- Refer to the included Clementine algorithm user manual, which explains the theories and methods behind the algorithms offered in Clementine

Evaluation

- Easily evaluate models using lift, gains, profit, and response graphs
  - Use a one-step process that shortens project time when evaluating multiple models
  - Define hit conditions and scoring expressions to interpret model performance
- Analyze overall model accuracy with coincidence matrices and other automatic evaluation tools

Deployment

Clementine offers a broad array of deployment capabilities to meet your organization's needs. Models built in Clementine can be directly deployed into other SPSS predictive applications as well as in other vendors' technologies.

- Clementine Solution Publisher™ (optional**)
  - Automate the export of all operations, including data access, data manipulation, text mining, model scoring (including combinations of models) and post-processing
  - Use a runtime environment for executing image files on target platforms
- PredictiveCallCenter
  - Automatically export Clementine streams for use in PredictiveCallCenter to make real-time customer recommendations
  - Combine exported Clementine streams with PredictiveCallCenter models, business rules, and exclusions to optimize customer interactions
- Cleo™ (optional**)
  - Implement a Web-based solution for rapid model deployment
  - Enable multiple users to simultaneously access and immediately score single records, multiple records, or an entire database, through a customizable browser-based interface
- Clementine Batch
  - Automate production tasks while working outside the user interface
    - Automate Clementine processes from other applications or scheduling systems
    - Generate encoded passwords
    - Call Clementine processes via the command line
- Scripting
  - Automate command-line scripts or scripts associated with Clementine streams to automate repetitive tasks in the user interface. Scripts generally perform the same types of actions that you otherwise would carry out using a mouse or keyboard.
    - Execute selected lines from a stream, SuperNode, or stand-alone script using an icon on the toolbar
    - Update stream parameters within a stand-alone script
- Export generated models as PMML 2.1
  - Perform in-database scoring, which eliminates the need for—and costs associated with—transferring data to client machines or performing calculations there
  - Deploy Clementine PMML models to IBM DB2 Intelligent Miner Visualization and Intelligent Miner Scoring
- Use the bulk-loading capabilities of your database
  - Increase performance during data export by using your database's bulk loader. Fine-tune various options, including row-wise or column-wise binding for loading via ODBC, and batch-size settings for batch commits to the database.

SPSS Model Manager

- Centralize data mining projects to leverage organizational knowledge
  - Save streams, models, and other objects in a central, searchable repository
  - Group streams in folders and secure folders and streams by user or user groups
  - Provide permission-based access to protect privacy of sensitive information
  - Reuse the most effective streams and models to improve processes and increase the accuracy of results
  - Search on input variables, target variables, model types, notes, keywords, authors, and other types of metadata
- Ensure reliable results by controlling versions of predictive models
  - Automatically assign versions to streams and other objects; Protect streams from being overwritten through automatic versioning

Scalability

- Use in-database mining to leverage parallel database implementations
- Use in-database modeling to build models in the database using leading database technologies
- Minimize network traffic via intelligent field projection, which means that Clementine pulls data only as needed from your data warehouse and passes only relevant results to the client

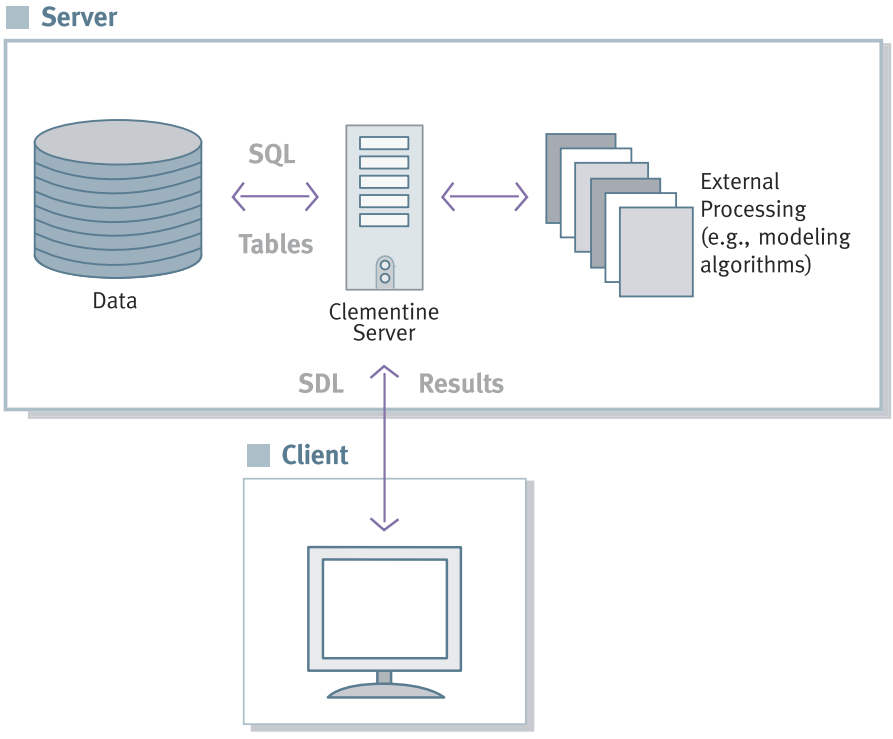
System requirements

Clementine Client

- Operating system: Microsoft Windows XP Home Edition, Windows XP Professional, Windows 2000 Professional
- Hardware:
  - Intel® Pentium®-compatible processor or faster
  - Monitor: XGA monitor with 1024 x 768 resolution or higher recommended
  - Memory: 512MB RAM recommended
  - Minimum free disk space: 320MB
  - A CD-ROM drive is required for installation
- Software: Microsoft Internet Explorer 6.0 or later for running the help system. Installing Clementine installs the Java Virtual Machine: Sun Java Runtime Environment 1.4.1_02.
- For modeling with Microsoft Decision Trees:
  - Clementine Client running in local mode or against a Clementine Server installation on Windows
  - Microsoft SQL Server with Microsoft Analysis Services (Service Pack 3 or higher)
- For modeling with Oracle Data Mining:
  - Clementine Client running in local mode or against a Clementine server installation on Windows or UNIX®
  - Oracle 10*g* with Oracle Data Mining installed
- For modeling with IBM Intelligent Miner:
  - Clementine Client running in local mode or against a Clementine Server installation on Windows or UNIX
  - IBM DB2 Enterprise Edition 8.2 with Intelligent Miner version 8.2. The Intelligent Miner Visualization tool is also supported as an optional add-on.

Clementine Server, Clementine Solution Publisher Runtime, and Clementine Batch

- Operating system: Windows Server™ 2003 or 2000; Sun™ Solaris™ 8 or 9, with 32-bit support; 64-bit support on Solaris 9 (SPARC 64-bit machine) or Solaris 10. HP-UX 11i; IBM AIX® 4.3.3 or AIX 5L, version 5.1 or higher; or OS/400® (on the IBM eServer iSeries™) V5R2 with OS/400 Portable Applications Solution Environment (PASE, 5722-SS1 Option 33)
- Hardware:
  - Pentium-compatible processor if running on Windows; UltraSPARC II or better for Solaris; PA-RISC processor and HP Workstation for HP/UX; PowerPC processor, 233MHz or faster, and IBM RS/6000® for AIX; or IBM iSeries server for OS/400
  - Memory: 512MB RAM minimum
  - Minimum free drive space: 128MB of available disk space are required for installation. Additional free disk space is required to run the program (for temporary files). 1GB is recommended.
  - For Clementine Solution Publisher Runtime, the minimum free disk space required to install the software is 64MB, plus at least twice the disk space of the amount of data to be processed
  - A network adapter running TCP/IP protocol
  - A CD-ROM drive is required for installation
- Software:
  - Clementine Client software must be at the same release level as the Clementine Server software
  - For AIX installations, the Visual Age C++ runtime is required
  - For HP-UX installations, C++ runtime libraries must be installed



Clementine provides data mining scalability by using a three-tiered architecture, as shown in this diagram. The Clementine Client tier (shown at the bottom) passes stream description language (SDL) to Clementine Server. Clementine Server then analyzes particular tasks to determine which it can push to the database. After the database runs the tasks that it can process, it passes only the relevant aggregated tables to Clementine Server. If you are using a CEMI, Clementine Server passes the relevant tasks to that particular external process.

** Features are subject to change based on the final product release* ☐ *Symbol indicates a new feature* ***Separately priced modules*

Clementine Application Templates

- Clementine 7.2 or later
- Memory: 512MB RAM recommended

Cleo

- Web server: Cleo requires at least one server computer that meets the following minimum requirements. Using additional processors, faster processors, and more RAM will improve performance.
  - Operating system: Windows Server 2003 or 2000, Windows NT® 4.0 Server with Service Pack 5 or higher (cannot be installed on Windows NT Terminal Server), or Solaris 7 or later
  - Hardware: Pentium-compatible processor, 500MHz or faster, if running on Windows; UltraSPARC II or better for Solaris
  - Memory: 512MB RAM
  - Minimum free drive space: 700MB of available disk space
  - A graphics adapter with 800 x 600 resolution (SVGA) or higher, capable of displaying at least 256 colors
  - A network adapter running the TCP/IP protocol
- Repository: the system requires a database to serve as a repository for published content, framework settings, and other information. The following databases are supported:
  - Microsoft SQL Server 2000
  - Oracle 8*i*, version 8.1.7
- Data warehouse: the system can be configured to access data from a data warehouse or database. The system has only been tested with SQL Server 2000 and Oracle 8*i* databases.
- Web client: content is delivered to clients as standard HTML pages. Supported browsers include:
  - Internet Explorer version 5.5 with Service Pack 2 or version 6.0 for Windows
  - Internet Explorer version 5.2 for Macintosh®
  - Netscape 6.2

Text Mining for Clementine***

Client version requirements:

- Clementine 9.0 or later
- Operating system: Windows XP Professional, Windows 2000 Professional
- Memory: 512MB RAM
- Minimum free disk space: 85MB, plus space for databases
- Web browser: Internet Explorer 5.0 or later or Netscape 6.0 or later is required to use the Viewer node

Server version requirements:

- Operating system: Windows Server 2003 or 2000, or Solaris 8 or 9. Note: Support on Solaris is available only for users of the 32-bit version of Clementine Server.
- Hardware: Pentium III processor, 1GHz or faster if running on Windows, or Sun UltraSPARC II or better if running on Solaris
- Memory: 512MB RAM
- Minimum free disk space: 85MB, plus space for databases

Web Mining for Clementine 1.1

Client version requirements:

- Clementine 8.0 or later
- Operating system: Windows XP Home Edition, Windows XP Professional, or Windows 2000 Professional with Service Pack 2 or later
- Memory: 512MB RAM
- Minimum free disk space: two times the amount of raw Web data being processed
- Software: Excel 2000 for events configuration

Server version requirements:

- Operating system: Windows XP Home Edition, Windows XP Professional, Windows 2000 Professional with Service Pack 2 or later, or Windows Server 2003 or 2000
- Memory: 512MB RAM
- Minimum free disk space: twice the amount of raw Web data being processed
- Optional database: SQL Server 2000

** Features are subject to change based on the final product release **Separately priced modules ***Except for Japanese-language version. Those requirements can be found at* ***www.spss.com/lexiquest/systemrequirements***

SPSS®

To learn more, please visit www.spss.com. For SPSS office locations and telephone numbers, go to www.spss.com/worldwide.

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners.
© 2004 SPSS Inc. CLM9SPC-1104